UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

February 28, 2001

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES B-11*

MEMORANDUM FOR Howard Hogan
Chief, Decennial Statistical Studies Division

From: Donna Kostanich DK
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared by: Michael Starsinic, Dennis Sissel, and Mark Asiala
Variance Estimation Team

Subject: Accuracy and Coverage Evaluation: Variance Estimates by Size of Geographic Area

The attached document was prepared, per your request, to assist the Executive Steering Committee on A.C.E. Policy in assessing the data with and without statistical correction.

This report focuses mainly on distributions of coefficients of variation of A.C.E. small-area estimates. In addition to summarizing results for selected subnational geographic areas, it also, where applicable, compares them to results from the 1990 Post-Enumeration Survey.

DSSD Census 2000 Procedures and Operations Memorandum Series B-11*
February 28, 2001

Accuracy and Coverage Evaluation: Variance Estimates by Size of Geographic Area

Michael D. Starsinic, Charles D. Sissel, and Mark E. Asiala
U.S. Census Bureau

TABLE OF CONTENTS

Appendix

Accuracy and Coverage Evaluation 2000: Variance Estimates by Size of Geographic Area

prepared by Michael D. Starsinic, Charles D. Sissel, and Mark E. Asiala

Executive Summary

What were our variance expectations in the 2000 A.C.E.?

A useful measure of an estimate's "reliability" is the coefficient of variation (CV), defined as the ratio of the square root of the variance of an estimate and the expected value of the estimate: the smaller the value of the CV, the more "precise" the estimator is.

We expected the CVs for the 2000 A.C.E. to be lower than the corresponding 1990 PES CVs. We investigated four geographic areas - states, congressional districts, places over 100,000 census population, and counties over 100,000 census population. We expected lower CVs because

- The housing unit sample size for the A.C.E. was almost double that of the PES (300,913 versus approximately 165,000).
- Better measures of population size were available during sample selection of clusters.
- Sampling weights were less variable.

These improvements did lead to much smaller sampling variances, as expected. The actual reduction was larger than the 25% reduction that would be expected from the increase in sample size alone.

How do the CVs compare between the 2000 A.C.E. and the 1990 PES at the state level?

- There was about a 40 percent decrease in the median CV, from 0.406 percent to 0.240 percent.

- The maximum observed CV decreased from 0.933 percent to 0.804 percent.

- Forty-seven states saw their CVs decline, with an average reduction of 36.8 percent.

- The expectation that "most [state-level] CVs will be less than 0.5 [percent]" (Kostanich, 1999) was met.

At the congressional district level?

- There was about a 40 percent decrease in the median CV, from 0.499 percent to 0.297 percent.

- The maximum observed CV decreased from 2.007 percent to 0.948 percent.

At the place level?

- There was about a 50 percent decrease in the median CV, from 0.629 percent to 0.314 percent, for places with a census population greater than 100,000.

- The maximum observed CV decreased from 1.435 percent to 1.702 percent.

At the county level?

- There was about a 40 percent decrease in the median CV, from 0.510 percent to 0.310 percent, for counties with a census population greater than 100,000.

Introduction and Background

The sampling variance of the A.C.E. is expected to be smaller than the sampling variance of the 1990 Post-Enumeration Survey (PES). Several differences between the A.C.E. and the PES lead us to this conclusion.

- The much larger sample size of the A.C.E. as opposed to the PES: 300,913 housing units in 11,303 clusters for the A.C.E. (Fenstermaker, 2000), versus approximately 165,000 housing units in approximately 5,000 clusters for the PES (Obenski & Fay, 2000). In general, a sample size increase of this magnitude will decrease the sampling variance.

- Better measures of population size in the sample selection of block clusters.

- Reduction in the variability of sampling weights.

Also, a 20 percent targeted sample of A.C.E.-selected clusters (Targeted Extended Search or TES) was chosen and searched for additional matches and duplicates in blocks surrounding sampled clusters. This should reduce the variance of the DSE as opposed to not searching surrounding blocks at all. However, the TES will increase the sampling variance of the A.C.E. relative to the 1990 PES, since the TES was a sample-based operation and a similar operation in the PES was implemented in all block clusters, thus creating no additional sampling variance.

However, the implementation of a more efficient operation relative to the 1990 PES can result in a reduction of systematic error in the estimates. Many of these additional matches and duplicates are highly clustered in a relatively small number of blocks. Because the TES is targeted to clusters with a large number of non-matched or mis-geocoded housing units (which are good indicators of additional matches and duplicates), it will be more efficient than the PES surrounding block search in finding additional matches and duplicates. (Navarro, 2000)

Variance Estimation Methodology

The A.C.E. survey was a multi-phase sample, which increased the difficulties of estimating the sampling variance. Multi-phase sampling differs from multi-stage in the following way: in a multi-stage design, the information needed to draw all stages of the sample is known before the sampling begins; in a multi-phase design, the information needed to draw the nth phase of the sample is unobtainable until the n-1st phase of the sample is completed. A methodology based in part on the Rao-Shao jackknife variance estimator (Rao & Shao, 1992) takes into account the multi-phase nature of the A.C.E. The estimation of the variance due to the A.C.E. attempts to capture these components of the variance (the relative contribution to the sampling error from the components is not considered in this analysis):

- Sampling variance due to the initial Listing sample.
- Sampling variance due to the A.C.E. Reduction and Small Block Subsampling.
- Sampling variance due to the Targeted Extended Search (TES) sample.
- Variance due to the imputation of correct enumeration, match, and residence probabilities for unresolved cases.

This estimate of variance is only intended to include the error from the above four components, and is not intended to quantify nonsampling errors, other than the error due to imputed probabilities given the imputation model. Components of sampling error which are not incorporated into the variance estimates are the error due to weight trimming, and the error due to large block subsampling.

One component of the nonsampling error which we have not incorporated is the "synthetic" (or model) error. This model assumes that the coverage rate is uniform over all areas within post-strata. To the extent that areas deviate from this assumption we are introducing synthetic or model error. The accuracy of this methodology may decrease in areas where localized effects not reflected in the post-stratification affect the true sampling variance. This discrepancy becomes more pronounced as the population of an area decreases. Thus, caution should be used in comparisons between areas of different sizes.

This new methodology directly estimates variances only for the final collapsed post-strata. We compute all other variances using a variance-covariance matrix for the post-stratum coverage correction factors (CCFs), which is the major output of the variance estimation process. The estimated ("synthetic") variance of any population estimate can be computed using this matrix and the unadjusted census counts, broken down by post-stratum and excluding persons out-of-scope of the A.C.E. (For more information see Kim et al (2000) and Starsinic & Kim (2001).)

$$\begin{aligned}
\hat{X}_s &= \text{Synthetic household population estimate for geographic area s} \\
&= \sum_{\text{post-strata h}} \hat{X}_{sh} \\
&= \sum_{h=1}^{416} C_{sh} \times CCF_h \text{, where } C_{sh} = \text{Census count of post-stratum h in geographic area s}
\end{aligned}$$

These are preliminary estimates of the synthetic household population. The final estimates will have undergone a controlled rounding procedure to produce integer counts of persons. These final controlled rounded values will differ very little from the preliminary estimate derived from the CCFs, so their use has almost no impact on the estimates and no effect on our conclusions.

$$\begin{aligned}
\mathrm{Var}(\hat{X}_s) &= \text{synthetic variance for synthetic household population estimate } \hat{X}_s \\
&= \mathrm{Var}\Big(\sum_{h=1}^{416} \hat{X}_{sh}\Big) \\
&= \sum_{h=1}^{416}\sum_{h'=1}^{416} \mathrm{Cov}(\hat{X}_{sh}, \hat{X}_{sh'}) \\
&= \sum_{h=1}^{416}\sum_{h'=1}^{416} \mathrm{Cov}(C_{sh} \times CCF_h, C_{sh'} \times CCF_{h'}) \\
&= \sum_{h=1}^{416}\sum_{h'=1}^{416} C_{sh} \times C_{sh'} \times \mathrm{Cov}(CCF_h, CCF_{h'})
\end{aligned}$$

Computing measures of reliability

Variances are useful when associated with an estimate but lose their utility when taken out of context. A variance of 5 implies different things for an estimate of 500 as opposed to an estimate of 0.5. A useful measure of an estimate's "reliability" is the coefficient of variation (CV), defined as the ratio of the square root of the variance of an estimate and the expected value of the estimate: the smaller the value of the CV, the more "precise" the estimator is. We will concentrate on the CV of the synthetic (small-area) estimate.

For any desired population estimate, geographic or otherwise:

$$\text{Synthetic total population estimate} = \text{Synthetic household population estimate } (\hat{X}) + \text{“Residual” count}$$

where the "Residual" count are persons out-of-scope of the A.C.E. sample. These include institutionalized and non-institutionalized group quarters persons, persons counted in Service Based Enumeration (SBE), and persons enumerated in the Remote Alaska operation.

The CV is computed as:

$$CV = \frac{\sqrt{\mathrm{Var}(\text{Synthetic total population estimate})}}{\text{Synthetic total population estimate}}$$

Since the Residual population is excluded from the A.C.E. sample, it adds no sampling variance, and the variance of the synthetic estimate is the same as the variance of the corresponding A.C.E estimate described above.

Comparing Sampling Errors for the A.C.E. And the PES

Overall, as expected, the CVs for the A.C.E. were smaller, for corresponding geographies, than the CVs for the PES. The median CVs for the A.C.E. were at least 40 percent smaller than those for the PES. In fact, all of the distributional statistics (mean, median, minimum, maximum, and first and third quartiles) were lower for the A.C.E. for all geographies, with only one small exception at the county level. The actual reduction was larger than the 25% reduction that would be expected from the increase in sample size alone. (Farber, 2001)

Table 1 provides national summary statistics from both the 2000 synthetic estimates and the 1990 PES estimates.

- Estimates for four geographic levels: states, congressional districts (103rd for 1990, 106th for 2000), places with a Census population greater than 100,000 (determined separately on 1990 and 2000 data), and counties with a Census population greater than 100,000 (determined separately on 1990 and 2000 data).

- Standard descriptive statistics for the CVs: number of observations (within each geographic area), mean, minimum, maximum, median, and first and third quartiles.

- Other statistics: mean size (in population) of the geographic area, and Margin of Error, which denotes the 90-percent margin of error for a geographical area of size equal to the mean size with a CV equal to the mean CV.

$$\text{Margin of Error} \doteq 1.645 \times \text{Synthetic Population Estimate} \times \text{CV}$$

The 1990 PES data in the table are based on the final 357 post-strata definitions, and come from Thompson (1992).

Graphs 1-4 show the distribution of the CVs for the four sets of geographic areas.

- The "a" series of graphs at the top of each page gives the distribution for the 2000 A.C.E. CVs, while the "b" series at the bottom gives the distribution for the 1990 PES CVs. The scales along the x-axis of each pair of graphs is identical to make it easier to compare the distributions between the A.C.E. and the PES.

- Graph 1c shows the distribution of the percent difference between the state CV estimates between the PES and the A.C.E.

- Graphs 3c and 4c show the frequency and average CV for all places and counties, not just those with a population of 100,000 or greater, for both the PES and the A.C.E. The frequencies are represented by the vertical bars and the average CVs by the lines.

How do the CVs compare between the 2000 A.C.E. and the 1990 PES at the state level?

- At the state level, the median CV decreased from 0.406 percent to 0.240 percent, about a 40 percent decline.

- Forty-seven states saw their CVs decline, with an average reduction of 36.8 percent.

- Four states, however, saw their CVs increase. All four of these states have small populations.

- The distribution in Graphs 1a and 1b clearly shows the trend of smaller CVs for the A.C.E.

- The maximum observed CV decreased from 0.933 percent to 0.804 percent.

- Two of the four outliers (0.6 percent and above) in 1a were also among the four outliers in 1b.

- From Graph 1c, three states had reductions in their CV of at least 60 percent.

At the congressional district level?

- At the congressional district level, the median CV decreased from 0.499 percent to 0.297 percent, about a 40 percent decline.

- The trend in the distributions in Graphs 2a and 2b shows a similar trend as the results on states.

- The maximum observed CV decreased from 2.007 percent to 0.948 percent.

- The congressional district with the largest CV in the A.C.E. has a population distribution in which several post-strata with high CVs are prevalent. (Davis, 2001)

Although both the 103rd and 106th congressional district boundaries were based on the 1990 census, some boundary changes did occur. Therefore, direct comparison between 103rd and 106th congressional districts was not attempted.

At the place level?

- At the place level, the median CV decreased from 0.629 percent to 0.314 percent, about a 50 percent decline.

- The trend to lower CVs in the A.C.E. as shown in Graphs 3a and 3b is striking, with more than 80 percent of the places for the A.C.E. falling into a CV range (0.2 percent to 0.4 percent) which contained only three places in the PES.

- The maximum observed CV decreased from 1.702 percent to 1.435 percent.

- The two outlying places in Graph 3a both had large concentrations of high-CV post-strata.

- Graph 3c shows a generally decreasing trend for CVs for places as the population increases. The average CV for each of the 10 population ranges in the graph is smaller in the A.C.E. than the PES.

- The number of places with a population of 100,000 or more increased from 195 to 245. The mean size dropped because many of the places that met the cutoff in 2000 but not in 1990 had relatively smaller populations.

The cutoff of 100,000 was applied separately to the PES and the A.C.E. data, and was based on the census population for the respective census. Direct comparisons between PES and A.C.E. CVs of specific places was not attempted because different definitions of "place" were used in the respective calculations - census place for the PES and FIPS place for the A.C.E. This may explain why the number of total places in 1990 is smaller than the total for 2000.

At the county level?

- At the county level, the median CV decreased from 0.510 percent to 0.310 percent, about a 40 percent decline.

- Once again, the distribution in Graph 4a is much more concentrated to the left (smaller CVs) than the distribution in Graph 4b.

- The county with the largest CV (1.498 percent) in the A.C.E. actually had a larger CV than any county in the PES (maximum 1.483 percent). This is the only time one of the six distribution statistics was larger for the A.C.E. than the PES for any of the geographic areas we investigated.

- The county with the highest CV in the A.C.E. had the second highest CV in the PES, and contains the place with the highest CV in the A.C.E.

- Graph 4c shows a generally decreasing trend for CVs for counties as the population increases. The average CV for each of the 10 population ranges in the graph is smaller in the A.C.E. than the PES.

- In the PES, the average CV increased over the last two population ranges (the two largest), after reaching a minimum in the 100,000 to 499,999 range. In the A.C.E., the decreasing trend continued over the last two population ranges.

The cutoff of 100,000 was applied separately to the PES and the A.C.E. data, and was based on the census population for the respective census. Because some county boundaries changed between 1990 and 2000, no direct comparisons of CVs for specific counties were made.

References

Davis, Peter, “Accuracy and Coverage Evaluation Survey: Dual System Estimation Results”, DSSD Census 2000 Procedures and Operations Memorandum Series B-9, February 28, 2001.

Farber, Jim, “Accuracy and Coverage Evaluation Survey: Quality Indicators of Census 2000 and the Accuracy and Coverage Evaluation”, DSSD Census 2000 Procedures and Operations Memorandum Series B-2, February 28, 2001.

Fenstermaker, Deborah, “Accuracy and Coverage Evaluation Survey: Sample Design Summary”, DSSD Census 2000 Procedures and Operations Memorandum Series R-33, May 30, 2000.

Haines, Dawn, “Accuracy and Coverage Evaluation Survey: Computer Specifications for Person Synthetic Estimation (U.S.)”, DSSD Census 2000 Procedures and Operations Memorandum Series Q-30, July 27, 2000.

Kostanich, Donna, “Census 2000 Accuracy and Coverage Evaluation Survey: Sample Allocation and Poststratification Plans”, DSSD Census 2000 Procedures and Operations Memorandum Series R-2, March 25, 1999.

Kim, Jae Kwang, Alfredo Navarro, and Wayne Fuller, “Replication Variance Estimation for Multi-Phase Stratified Sampling”, Internal Census Bureau Memorandum, July 24, 2000.

Navarro, Alfredo, “Accuracy and Coverage Evaluation Survey: Targeted Extended Search Plans”, DSSD Census 2000 Procedures and Operations Memorandum Series Q-18, January 12, 2000.

Obenski, Sally and Robert Fay, “Analysis of C.A.P.E. Findings on PES Accuracy at Various Geographic Levels”, Internal Census Bureau Memorandum, June 6, 2000.

Rao, J.N.K., and Jun Shao, “Jackknife variance estimation with survey data under hot deck imputation”, Biometrika, 79, 811-812, 1992.

Starsinic, Michael, and Jae Kwang Kim, “Computer Specifications for Variance Estimation for Census 2000”, DSSD Census 2000 Procedures and Operations Memorandum Series V-5, February 22, 2001.

Thompson, John, “357 Post Enumeration Survey (PES) Estimates”. Internal Census Memorandum to the CAPE Committee and CAPE Working Group, April 24, 1992.

Table 1: US summary of distribution of CVs for population estimates by geographical area for 1990 PES and 2000 A.C.E.

Area	Source	Number	Mean Size	Mean	Margin of			Distribution of CVs		
				CV	Error*	Minimum	Q1	Median	Q3	Maximum
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
State **	A.C.E.	51	5,582,035	0.310%	28,506	0.159%	0.220%	0.240%	0.378%	0.804%
	PES	51	4,955,153	0.449%	36,623	0.322%	0.369%	0.406%	0.496%	0.933%
Congressional Districts ***	A.C.E.	435	653,103	0.330%	3,546	0.156%	0.250%	0.297%	0.375%	0.948%
	PES	435	579,567	0.557%	5,309	0.299%	0.420%	0.499%	0.628%	2.007%
Places > 100,000 ****	A.C.E.	245	315,037	0.343%	1,776	0.213%	0.283%	0.314%	0.361%	1.435%
	PES	195	335,637	0.673%	3,718	0.363%	0.536%	0.629%	0.747%	1.702%
Counties > 100,000 ****	A.C.E.	524	409,345	0.368%	2,481	0.201%	0.274%	0.310%	0.405%	1.498%
	PES	458	400,593	0.534%	3,519	0.285%	0.432%	0.510%	0.591%	1.483%

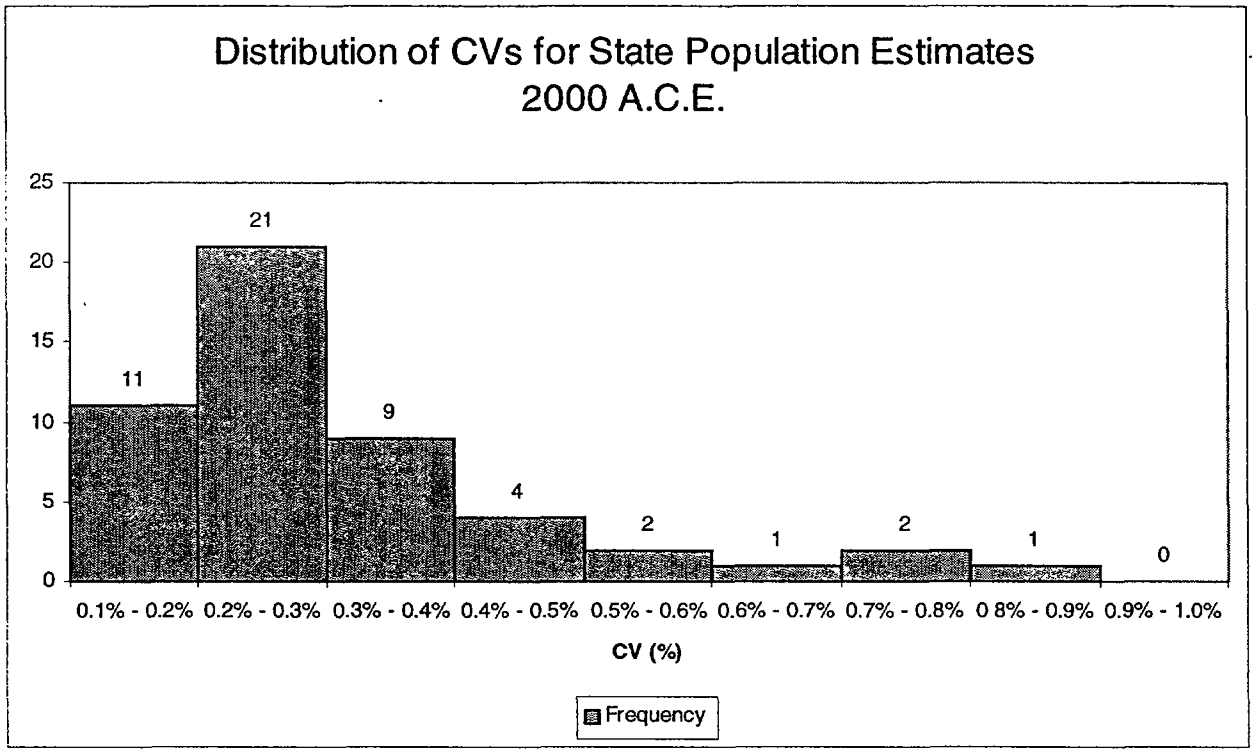
* - Margin of Error is calculated as 1.645 × standard error of the population estimate.

** - "State" includes all 50 states and the District of Columbia, but does not include Puerto Rico.

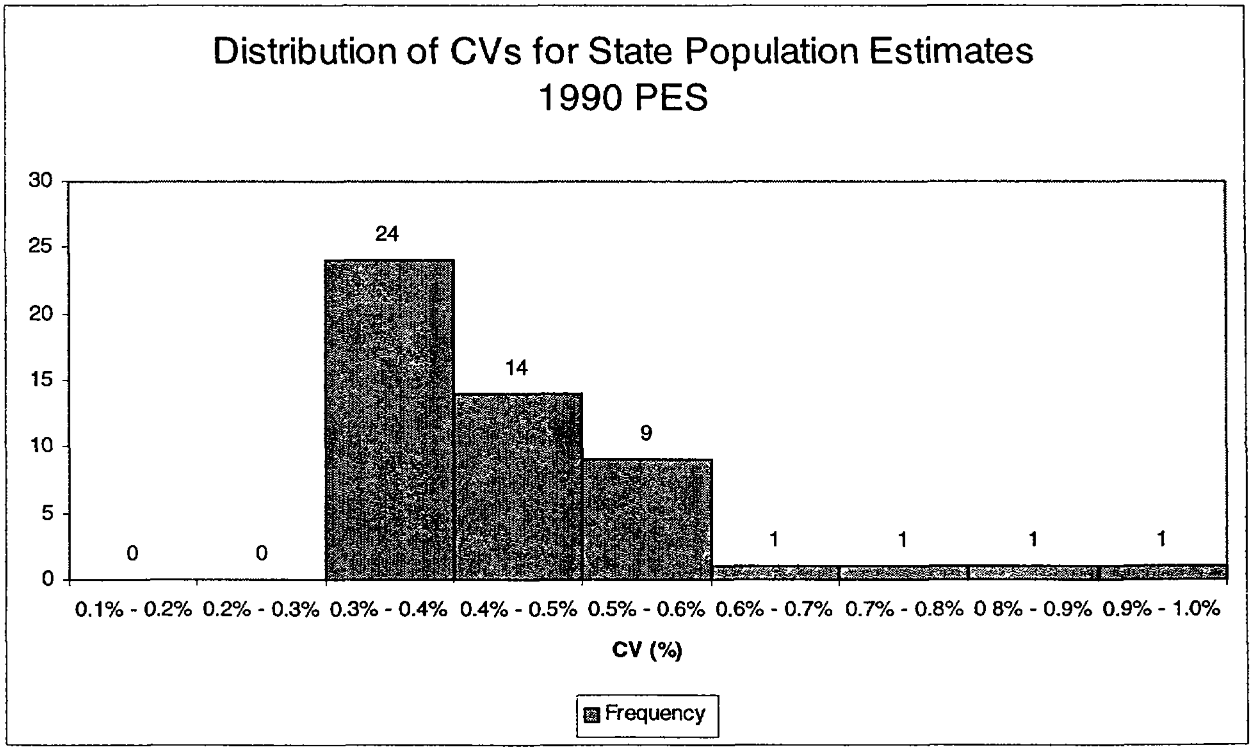
*** - 103rd Congressional Districts for the PES; 106th Congressional Districts for the A.C.E. Does not include the District of Columbia or Puerto Rico.

**** - Counties and places with census counts of more than 100,000 in the respective censuses, 2000 for A.C.E. and 1990 for PES.

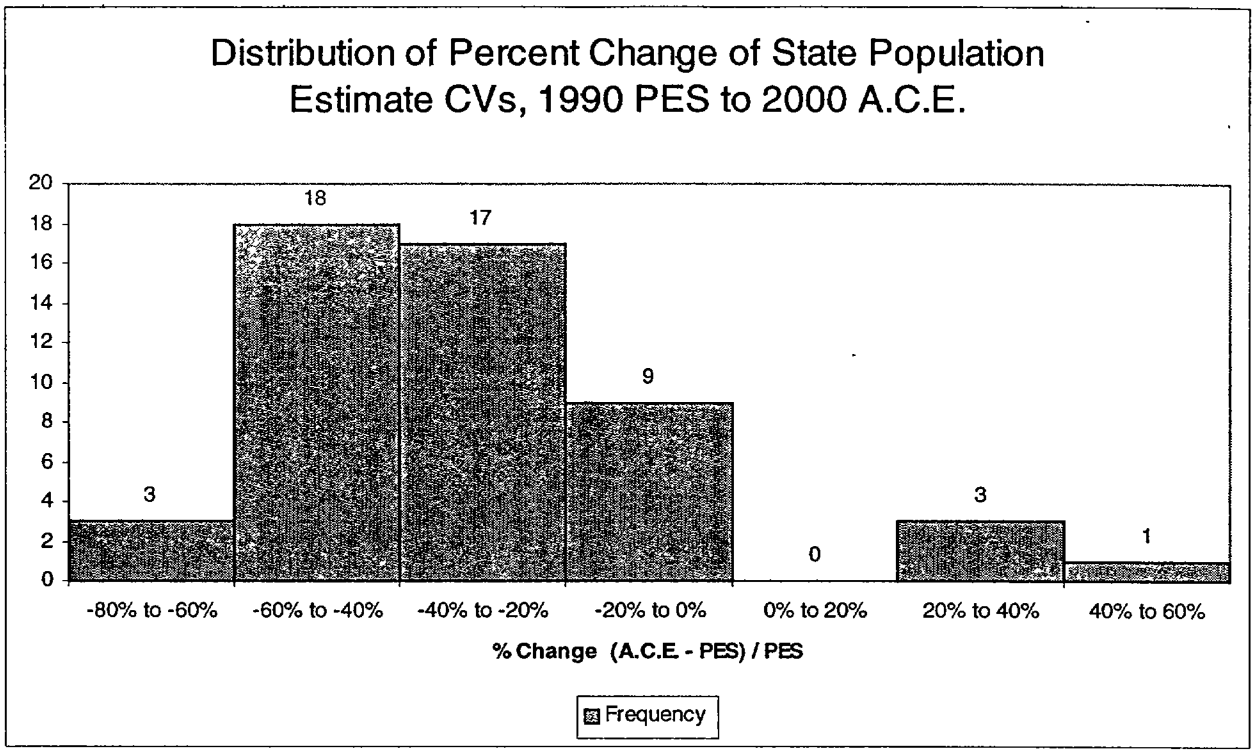
Graph 1a:



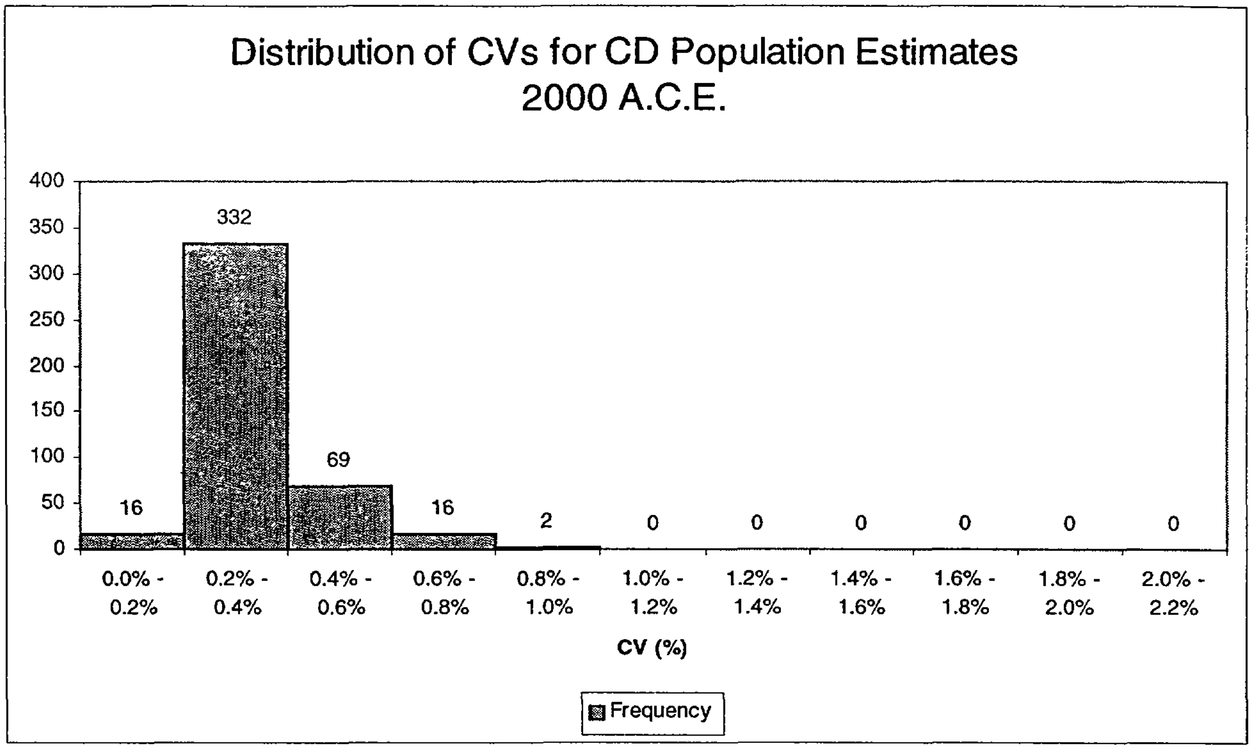
Graph 1b:



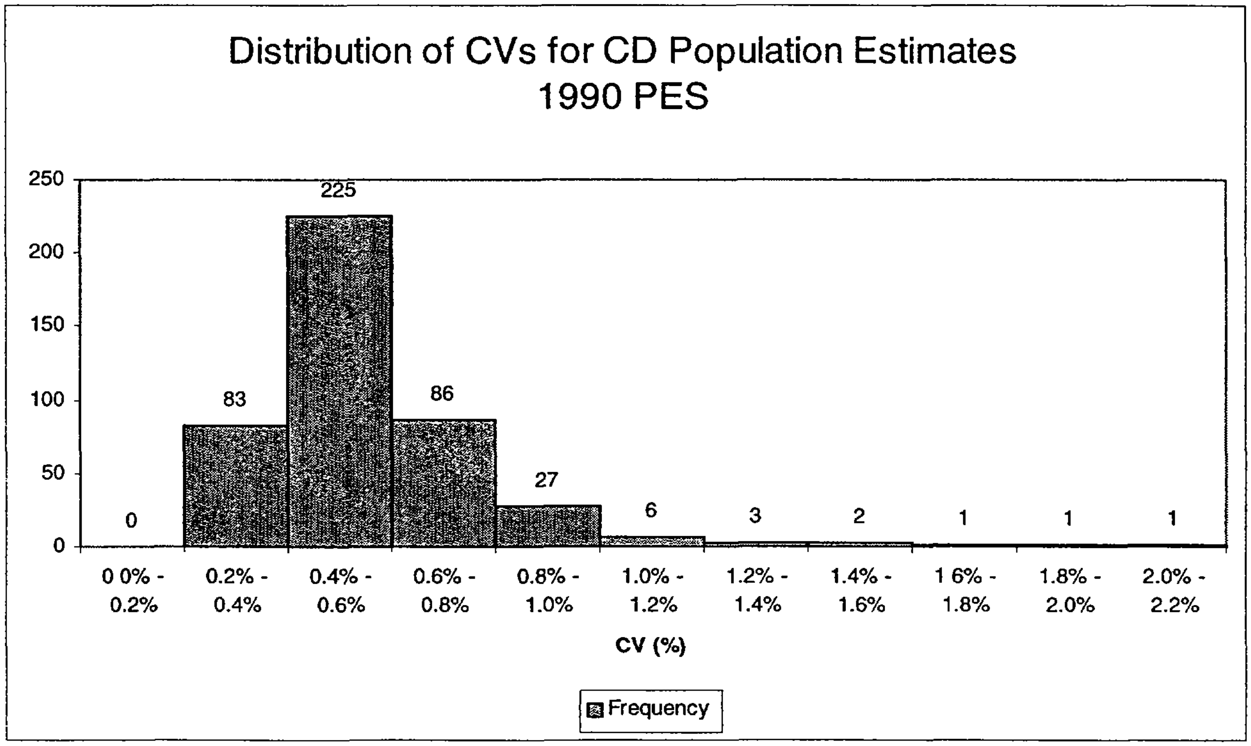
Graph 1c:


Distribution of Percent Change of State Population Estimate CVs, 1990 PES to 2000 A.C.E.
20
18
16
14
12
10
8
6
4
2
0
3
18
17
9
0
3
1
-80% to -60%
-60% to -40%
-40% to -20%
-20% to 0%
0% to 20%
20% to 40%
40% to 60%
% Change (A.C.E - PES) / PES
▨ Frequency

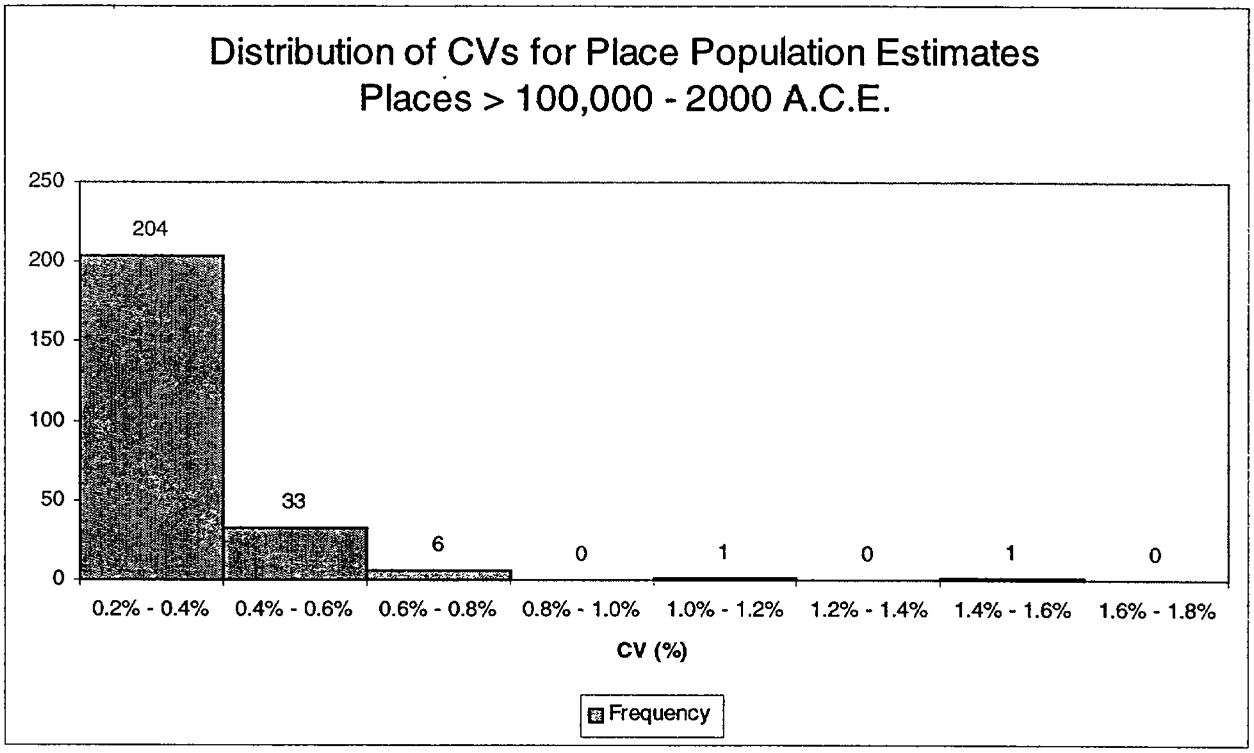
Graph 2a:



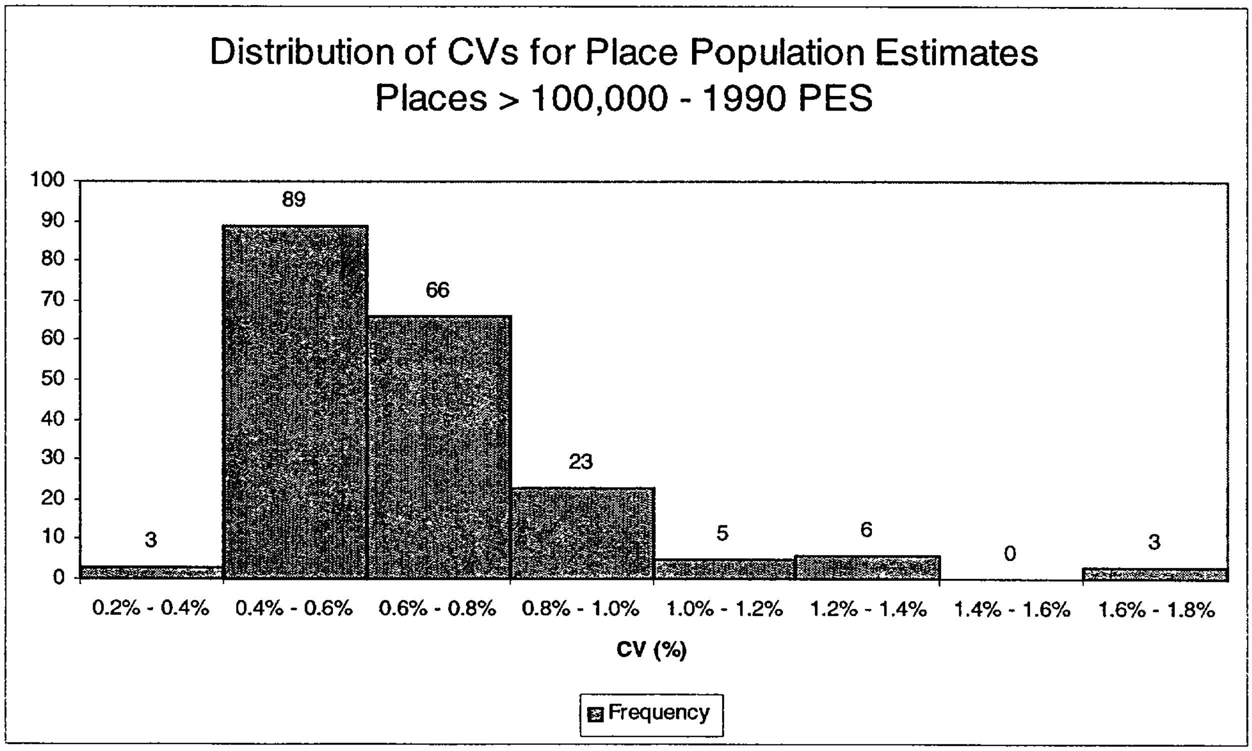
Graph 2b:



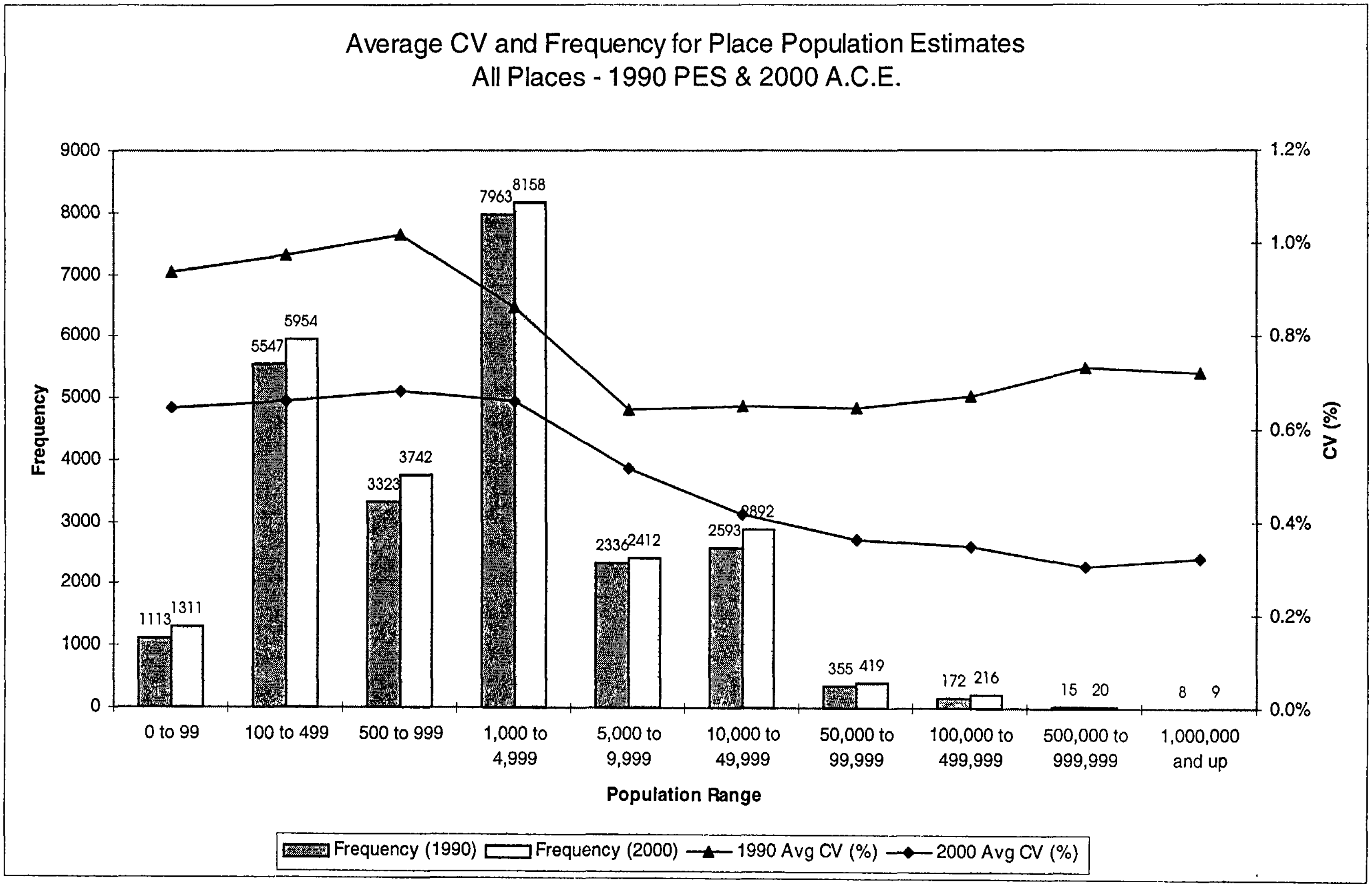
Graph 3a:



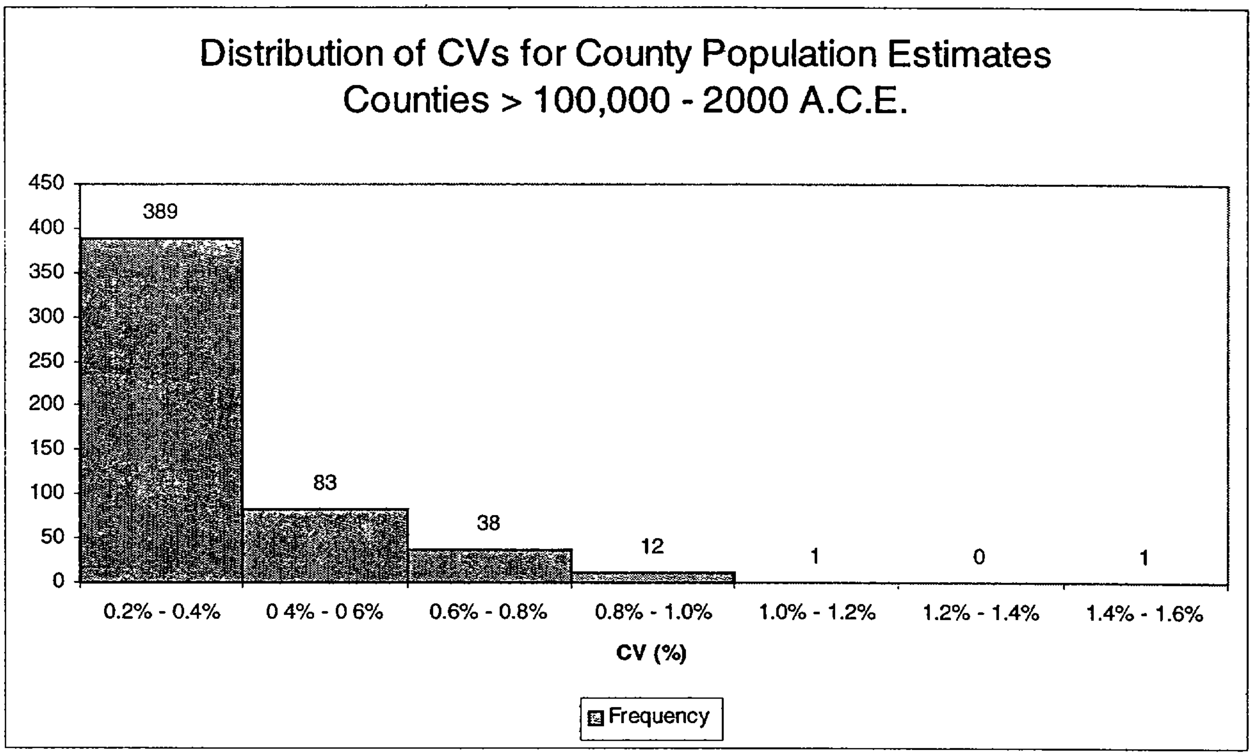
Graph 3b:



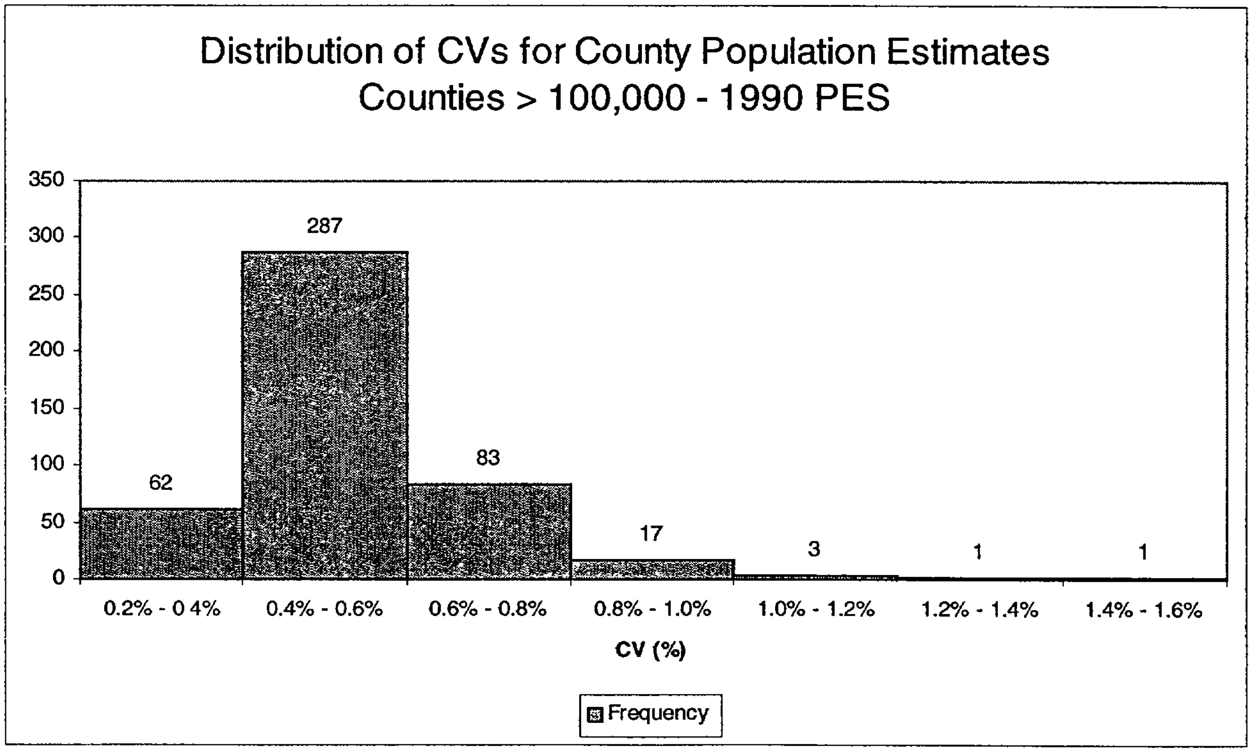
Graph 3c:


Average CV and Frequency for Place Population Estimates
All Places - 1990 PES & 2000 A.C.E.
Frequency
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
CV (%)
1.2%
1.0%
0.8%
0.6%
0.4%
0.2%
0.0%
1113
1311
5547
5954
3323
3742
7963
8158
2336
2412
2593
2892
355
419
172
216
15
20
8
9
0 to 99
100 to 499
500 to 999
1,000 to 4,999
5,000 to 9,999
10,000 to 49,999
50,000 to 99,999
100,000 to 499,999
500,000 to 999,999
1,000,000 and up
Population Range
Frequency (1990)
Frequency (2000)
1990 Avg CV (%)
2000 Avg CV (%)

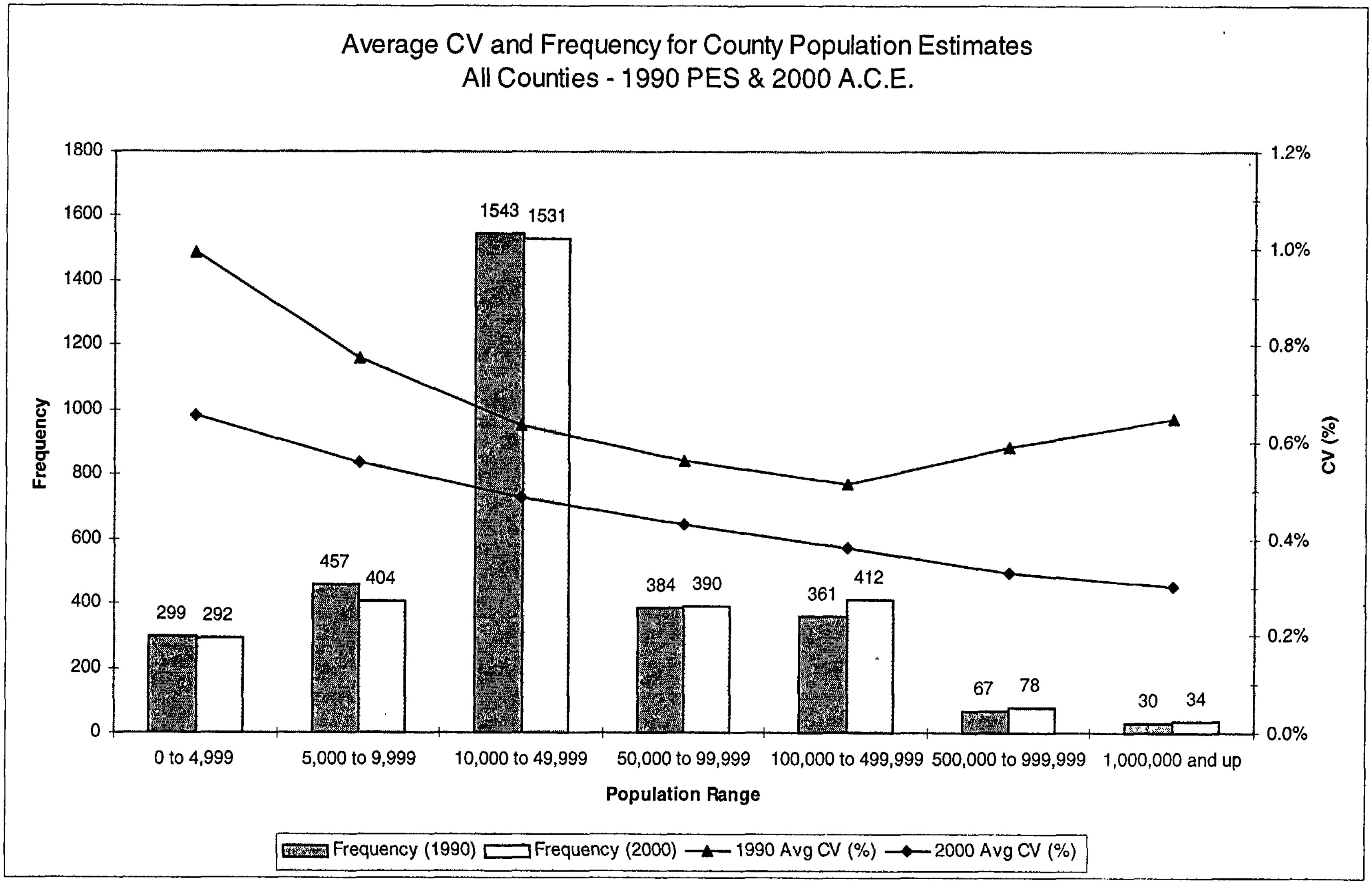
Graph 4a:



Graph 4b:



Graph 4c:



Average CV and Frequency for County Population Estimates
All Counties - 1990 PES & 2000 A.C.E.